Knowledge and Information Systems (2025) 67:5335–5354
https://doi.org/10.1007/s10115-025-02377-7

RESEARCH

CTVAE: Contrastive Tabular Variational Autoencoder for imbalance data

Alex X. Wang[1] · Minh Quang Le[2] · Huu-Thanh Duong[2] · Bay Nguyen Van[2] · Binh P. Nguyen[1]

Received: 5 November 2024 / Revised: 26 January 2025 / Accepted: 15 February 2025 /
Published online: 15 March 2025

Abstract

Class imbalance, where datasets often lack sufficient samples for minority classes, is a persistent challenge in machine learning. Existing solutions often generate synthetic data to mitigate this issue, but they typically struggle with complex data distributions, primarily because they focus on oversampling the minority class while neglecting the relationships with the majority class. To overcome these limitations, we propose the Contrastive Tabular Variational Autoencoder (CTVAE), which integrates conditional Variational Autoencoders with contrastive learning techniques. CTVAE excels at generating high-quality synthetic samples that capture the intricate data distributions of both minority and majority classes. Additionally, it can be seamlessly integrated with variants of the Synthetic Minority Oversampling Technique (SMOTE) for enhanced effectiveness. Experimental results demonstrate that CTVAE substantially improves classification performance on imbalanced datasets, offering a more robust and holistic solution to the class imbalance problem.

Keywords Imbalance data · Synthetic data · Deep learning · Contrastive learning · Data-centric AI

Alex X. Wang and Minh Quang Le have contributed equally to this work.

✉ Binh P. Nguyen
binh.p.nguyen@vuw.ac.nz

Alex X. Wang
alex.wang@vuw.ac.nz

Minh Quang Le
minh.le@ou.edu.vn

Huu-Thanh Duong
huu-thanh.duong@ou.edu.vn

Bay Nguyen Van
bay.van@ou.edu.vn

[1] School of Mathematics and Statistics, Victoria University of Wellington, Kelburn Parade, Wellington 6012, New Zealand

[2] Faculty of Information Technology, Ho Chi Minh City Open University, 97 Vo Van Tan, District 3, Ho Chi Minh City 70000, Vietnam

1 Introduction

Class imbalance in binary classification is a common challenge, where traditional machine learning (ML) methods often struggle to represent the minority class adequately [1]. This imbalance can lead to biased models and reduced performance, especially in high-stakes applications like medical diagnosis and defect detection [2, 3]. To address this, oversampling techniques have been developed to balance class distributions by creating synthetic instances. However, traditional methods like the Synthetic Minority Oversampling Technique (SMOTE) [4] generate synthetic data within the minority class region alone, without utilizing information from the majority class. This limitation can result in suboptimal outcomes, as it overlooks the complex interplay between classes and may fail to capture the broader data structure [5].

Class imbalance remains a critical challenge in machine learning, particularly when working with datasets where the minority class is significantly underrepresented. Recent advancements in deep learning-based generative methods have introduced powerful tools for addressing this issue. Models such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and diffusion models have shown promise in generating synthetic data to mitigate class imbalance [6, 7]. However, these methods often rely on unconditional generation [8], which tends to skew outputs toward the majority class, failing to capture the nuanced patterns of the minority class distribution. Furthermore, GANs are prone to mode collapse, and diffusion models, while effective, demand significant computational resources, limiting their application [9]. To overcome these challenges, we propose the Conditional Tabular Variational Autoencoder (CTVAE), a novel approach that integrates conditional generative modeling with contrastive learning to address class imbalance more effectively. Contrastive learning, a proven technique in tasks such as image classification and recommendation [10], enhances the model's ability to learn meaningful representations by bringing similar samples closer together and pushing dissimilar samples apart in the embedding space [11]. By leveraging this capability, CTVAE generates high-quality synthetic data that captures the distributions of both the majority and minority classes, ensuring a more balanced and accurate augmentation strategy. This approach not only addresses the limitations of existing generative models but also offers a computationally efficient solution, making it a practical choice for real-world imbalanced data scenarios.

CTVAE offers a robust advancement over existing oversampling methods, effectively addressing their limitations through the integration of majority and minority class information. Extensive experiments demonstrate CTVAE's consistent superiority over traditional approaches, largely due to its ability to generate diverse and representative synthetic samples that significantly enhance model performance, particularly in small, imbalanced datasets. By leveraging class information within advanced VAE architectures combined with contrastive learning, CTVAE creates high-quality synthetic data that better capture the complexities of underlying distributions. This integration not only boosts synthetic sample quality but also strengthens classification accuracy across a range of applications. Additionally, the strategic use of SMOTE and its variants further amplifies CTVAE's effectiveness in managing class imbalance. Empirical studies validate the efficacy of our method, highlighting its capability to improve model robustness and accuracy across various domains and datasets.

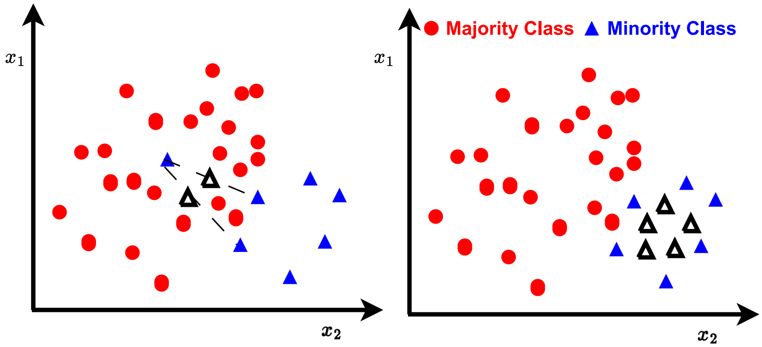

Fig. 1 Visual illustration of SMOTE and its variants issues

2 Related work

2.1 Imbalanced data problem

The class imbalance problem is a significant challenge in machine learning (ML) that affects various fields. It arises when the distribution of instances across classes is highly uneven, with the majority class substantially outnumbering the minority classes [12]. This imbalance often skews the training process [13], causing models to favor the majority class and overlook the minority class, resulting in biased predictions. Such biases can undermine the model's ability to accurately predict minority class instances and raise fairness concerns [14], particularly in real-world applications where equity and accuracy are critical. Addressing class imbalance is crucial to ensuring fair and reliable ML outcomes across domains [15]. To tackle this issue, previous research has focused on three main approaches: data level, algorithm level, and ensemble learning. Data-level methods, such as SMOTE, generate synthetic samples for the minority class (oversampling) or remove instances from the majority class (undersampling). Algorithm-level methods adjust learning algorithms by assigning different misclassification costs to classes to achieve balance [16]. Ensemble learning techniques, including bagging and boosting, improve performance by aggregating predictions from multiple models [17]. These strategies are explored in detail in comprehensive surveys on the topic [18].

2.2 SMOTE and its variants

SMOTE and its variants are widely used to address the class imbalance by generating synthetic samples for minority classes positioned between existing instances [19]. For instance, Borderline-SMOTE [20] targets boundary samples to enhance class separation, while ADASYN [21] focuses on generating more samples in areas with low minority class density. A more recent approach, SyMProD [22], leverages probabilistic distribution and nearest neighbors to generate synthetic minority instances, effectively addressing skewed datasets while reducing noise and class overlap. Hybrid approaches like SMOTE-Tomek and SMOTEENN combine oversampling and undersampling techniques to refine datasets and reduce noise [23]. However, these methods exhibit two critical issues despite their benefits, as illustrated in Fig. 1. First, they may overlook important information from the majority class, potentially introducing noise into the majority class space. Second, the synthetic instances they generate may lack the necessary diversity, failing to capture the complex distributions of the minority class adequately. Therefore, while these methods effectively increase minority class samples, they may not capture the variability crucial for improving classification performance, especially in highly imbalanced and complex datasets.

2.3 Conditional Variational Autoencoder

A Conditional Variational Autoencoder (CVAE) extends the VAE framework by incorporating additional conditioning information [24]. In a CVAE, the encoder network models the posterior distribution of the latent variable z given both the observed data x and conditioning information y as $q_\phi(z|x, y)$. This conditioning information y could represent any relevant auxiliary data, such as class labels or other contextual features, which helps the model generate data more accurately tailored to specific conditions. Similarly, the decoder network models the conditional distribution of the data given z and y as $p_\theta(x|z, y)$. By leveraging this conditional framework, the CVAE is capable of generating diverse data samples that reflect the desired conditions specified by y, making it particularly effective in applications where controlled generation is required. The objective of a CVAE is to maximize the Evidence Lower Bound (ELBO) on the log-likelihood of the observed data given y [25]:

$$\log p_\theta(x|y) \geq \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(x|z, y)] - \mathrm{KL}[q_\phi(z|x, y)||p(z|y)],$$

where the reconstruction loss $-\mathbb{E}_{q(z|x,y)}[\log p(x|z, y)]$ measures dissimilarity between input data x and its reconstruction given z and y. The regularization term $\mathrm{KL}[q(z|x, y)||p(z|y)]$ ensures the latent variable $q(z|x, y)$ aligns with a prior $p(z|y)$. By minimizing this loss, CVAE learns to reconstruct data while maintaining a latent space consistent with the prior conditioned on y. However, while ELBO is useful for unconditional generative modeling, it may not adequately capture the complex relationships between conditioned inputs and outputs [26]. To address this, contrastive learning enhances CVAE by focusing on discriminative representation learning, which involves comparing pairs of samples in the latent space [27].

2.4 Contrastive learning

Contrastive learning is a form of self-supervised learning where the model improves by contrasting pairs of data samples to bring similar ones closer and push dissimilar ones apart in a latent space [28]. This approach encourages the model to learn meaningful representations that capture the structure of the data [27]. Mathematically, contrastive learning is implemented using a contrastive loss function. Given a dataset of samples $x_1, x_2, \ldots, x_N$, for each anchor sample x_i, a positive sample x_j (from the same class or similar context) and a negative sample x_k (from a different class or context) are selected. The objective is to ensure that similar samples x_i and x_j have closer latent representations, while dissimilar samples x_i and x_k are represented farther apart.

Let f_θ denote a neural network parameterized by θ that maps data samples to a latent space. The contrastive loss function consists of two key components: a positive term and a negative term. The positive term encourages similar latent representations for similar samples, while the negative term encourages dissimilar latent representations for dissimilar samples. Here is the contrastive loss function applied to latent space vectors: The positive term encourages

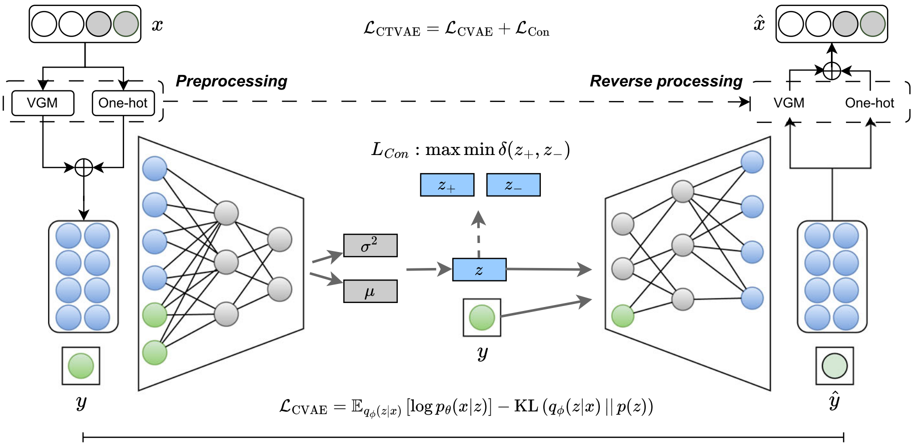
x
$\mathcal{L}_{\text{CTVAE}} = \mathcal{L}_{\text{CVAE}} + \mathcal{L}_{\text{Con}}$
$\hat{x}$
VGM
One-hot
Preprocessing
Reverse processing
VGM
One-hot
L_{Con} : max min $\delta(z_+, z_-)$
z_+
z_-
σ^2
μ
z
y
y
$\mathcal{L}_{\text{CVAE}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x) \,||\, p(z))$
$\hat{y}$

Fig. 2 Illustrating the CTVAE: During training, each data sample x undergoes preprocessing and is encoded along with label y to produce a latent space vector z. This vector, along with the label y, serves as input for the decoder, which engages in a reconstruction process incorporating the standard VAE loss and contrastive loss. The contrastive loss is introduced to enhance the model's ability to learn the latent space based on the label. During the generation phase, a randomly sampled latent space vector $\hat{z}$ and label y are fed into the trained decoder to generate synthetic data, followed by reverse processing to return the data to its original format, $\hat{x}$

similar samples z_i and z_j to have similar latent representations:

$$\text{pos\_loss}(z_i, z_j) = -\log \frac{\exp(f_\theta(z_i) \cdot f_\theta(z_j)/\tau)}{\sum_{k=1}^{N} \exp(f_\theta(z_i) \cdot f_\theta(z_k)/\tau)},$$

where τ is a temperature parameter that controls the smoothness of the distribution. The negative term encourages dissimilar samples z_i and z_k to have dissimilar latent representations:

$$\text{neg\_loss}(z_i, z_k) = -\log \frac{\sum_{k=1}^{N} \exp(f_\theta(z_i) \cdot f_\theta(z_k)/\tau)}{\sum_{k=1}^{N} \sum_{m=1}^{N} \exp(f_\theta(z_i) \cdot f_\theta(z_m)/\tau)}.$$

The overall contrastive loss function is the sum of the positive and negative terms:

$$\mathcal{L}_{\text{Con}}(z_i, z_j, z_k) = \text{pos\_loss}(z_i, z_j) + \text{neg\_loss}(z_i, z_k).$$

In this context, z_i, z_j, and z_k are latent space vectors obtained from the encoder of the CTVAE model, with z_j being a positive sample (similar to z_i) and z_k being a negative sample (dissimilar to z_i). By applying this loss function, the latent space is structured to reflect meaningful relationships within the data, ensuring that similar samples (e.g., from minority classes) are well represented. This contrastive framework is particularly impactful in scenarios with imbalanced datasets. By promoting tighter clustering of minority class samples in the latent space, the model enhances their representation [29].

3 Contrastive Tabular Variational Autoencoder

This section introduces CTVAE, our tabular data balancing model that integrates CVAE with contrastive learning. As shown in Fig. 2 and detailed in Algorithm 1, the process involves two main stages: (1) fine-tuning a CVAE to capture label-dependent feature interactions in

Algorithm 1 Contrastive Tabular Variational Autoencoder

```
**Input:** Real data samples $\{(x(i), y(i))\}_{i=1}^{m}$, learning rates $\alpha_{\text{enc}}, \alpha_{\text{dec}}$, contrastive loss coefficient $\beta$
**Initialize:** Autoencoder parameters $\phi, \theta$
**for** each training iteration **do**
    Sample $(x(i), y(i))$ from the training set
    Sample $z'(i)$ from the true prior $p_z$
    Sample $z$ from $q_\phi(z|x, y)$
    **Train the encoder/decoder** $(\phi, \theta)$**:**
    Compute the reconstruction loss:
    $L_{\text{CVAE}} \leftarrow ELBO$
    Compute the contrastive loss based on $y$:
        $positive\_pairs \leftarrow \|z_i - z_j\|_2$ for $y_i = y_j$
        $negative\_pairs \leftarrow \|z_i - z_j\|_2$ for $y_i \neq y_j$
    $L_{\text{Con}} \leftarrow \text{pos\_loss}(x_i, x_j) + \text{neg\_loss}(x_i, x_k)$
    Update encoder and decoder parameters:
        $\phi \leftarrow \phi - \alpha_{\text{enc}} \nabla_\phi (L_{\text{CAVE}} + \beta \cdot L_{\text{Con}})$
        $\theta \leftarrow \theta - \alpha_{\text{dec}} \nabla_\theta (L_{\text{CVAE}} + \beta \cdot L_{\text{Con}})$
**end for**
```

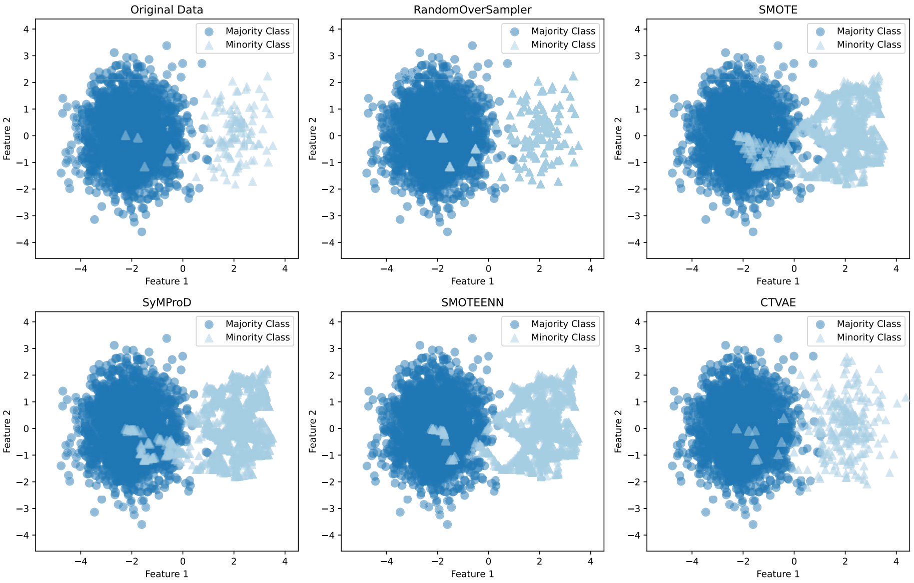

Fig. 3 Visual representation of the CTVAE and comparison with competing resampling algorithms on a simulated imbalanced dataset

the latent space and (2) using the trained decoder to generate synthetic data. We then provide detailed explanations of each component and summarize our approach.

3.1 Problem definition

The problem of data balancing in a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ as samples from a true data distribution $q(x|y)$ arises when there is a significant class imbalance between the minority $(y_i = 1)$ and majority $(y_i = 0)$ classes. The goal of a deep learning-based oversampling model is to build neural networks with parameters θ to describe a conditional distribution

$p_\theta(x|y)$, ensuring $p_\theta(x|y)$ matches $q(x|y)$ optimally during training. Using $p_\theta(x|y)$, one can generate more minority class samples to balance N_{neg} and N_{pos}.

3.2 Reversible data preprocessing pipeline

Our preprocessing pipeline, based on prior research [30], handles numerical and categorical columns. Continuous variables are processed using a variational Gaussian mixture model (VGM) [31] to handle non-Gaussian and multimodal distributions, normalized within each mode. Categorical features undergo one-hot encoding, converting unique elements into binary vectors. Ultimately, the processed numerical and categorical variables are concatenated for each row, denoted as r_j, making it ready for subsequent DGMs [32, 33]. The representation of the jth row is effectively captured by:

$$r_j = \alpha_{1,j} \oplus \beta_{1,j} \oplus \ldots \oplus \alpha_{n_c,j} \oplus \beta_{n_c,j} \oplus \gamma_j + d_{1,j} \oplus \ldots \oplus d_{n_d,j}.$$

3.3 Enhanced loss function

CTVAE is tailored for tabular data, leveraging class information to generate diverse synthetic minority samples. Contrastive learning complements this by promoting meaningful representations in the latent space, enhancing the quality and diversity of generated samples. Mathematically, the loss function of CTVAE is formulated as:

$$\mathcal{L}_{\text{CTVAE}} = \text{ELBO} + \lambda \mathcal{L}_{\text{Con}}(z_i, z_j, z_k),$$

where ELBO represents the Evidence Lower Bound and $\mathcal{L}_{\text{Con}}(z_i, z_j, z_k)$ denotes the contrastive loss function, with λ as a hyperparameter to adjust the strength of the regularization.

To validate our hypothesis on enhancing data balancing using CTVAE, we generated a synthetic binary dataset for analysis. As depicted in Fig. 3, we compare the original dataset with its oversampled version. The results clearly illustrate that traditional oversampling methods tend to introduce considerable noise in the majority sample area, often resulting in overlapping and redundant samples. Additionally, synthetic data produced by these conventional techniques typically cluster narrowly within the existing minority class area, which further reinforces the issues of limited representation and diversity in the minority class. This outcome supports our earlier discussions on the shortcomings of traditional approaches. In contrast, our proposed CTVAE leverages training on both majority and minority classes, effectively addressing these challenges. By integrating information from both classes, CTVAE reduces the noise present in the majority class area, creating a cleaner and more coherent dataset. Furthermore, it introduces a wider variety of synthetic samples into the minority class, leading to a more balanced distribution. This enhancement allows the model to capture the underlying patterns of the minority class more accurately, resulting in improved classification performance and robustness. Overall, the findings demonstrate the effectiveness of CTVAE in generating high-quality synthetic data that contribute to better data balancing and representation in imbalanced datasets.

Table 1 Details of the datasets used in this study

Abbr	Name	Source	#Rows	#Num	#Cat	IR
EC	ecoli	UCI	336	5	2	8.6
LM	libras_move	UCI	360	90	0	14.0
AR	arrhythmia	UCI	452	140	137	17.0
OI	oil	UCI	937	38	10	22.0
UC	us_crime	UCI	1994	98	1	12.0
YM	yeast_ml8	LIBSVM	2417	102	0	13.0
OZ	ozone_level	UCI	2536	71	0	34.0
SI	sick	UCI	3103	4	17	13.4
TH	thyroid	UCI	3772	6	15	15.0
AB	abalone	UCI	4177	6	3	9.7
WQ	wine_quality	UCI	4898	10	0	26.0
OD	optical_digits	UCI	5620	1	63	9.1
ST	statlog	UCI	6435	36	0	9.3
CO	coil_2000	UCI	9822	43	43	16.0
PD	pen_digits	UCI	10, 992	15	0	9.4
AD	adult	UCI	48,842	5	8	3.2

#Rows, #Num and #Cat represent the number of samples, numerical columns and categorical columns. IR stands for Imbalance Ratio

4 Experiments

4.1 Datasets

To evaluate our algorithm, we used 16 real-world public datasets of varying sizes, features, and distributions, commonly used in tabular model studies [30, 34, 35]. Detailed descriptions of each dataset and their key properties are summarized in Table 1. For multi-class datasets, we retained the minority class as-is and combined all other classes into a single majority class to create a binary classification setup. To prevent data leakage, datasets were split 80% for training and 20% for testing. All models were trained on the same training data with default hyperparameters. To ensure the reliability of our results, we repeated each experiment five times with different random seeds for data splitting and reported the average performance metrics [36].

4.2 Experimental settings

The experimental framework, shown in Fig. 4, investigates the use of conditional DGMs to improve classification performance on imbalanced datasets. We assess CTVAE against popular and effective resampling techniques, including SMOTE, SyMProD, and SMOTEENN, based on their proven performance in prior studies [22, 37]. The study includes five distinct experimental settings, each utilizing a unique data strategy alongside the original imbalanced data without any sampling applied:

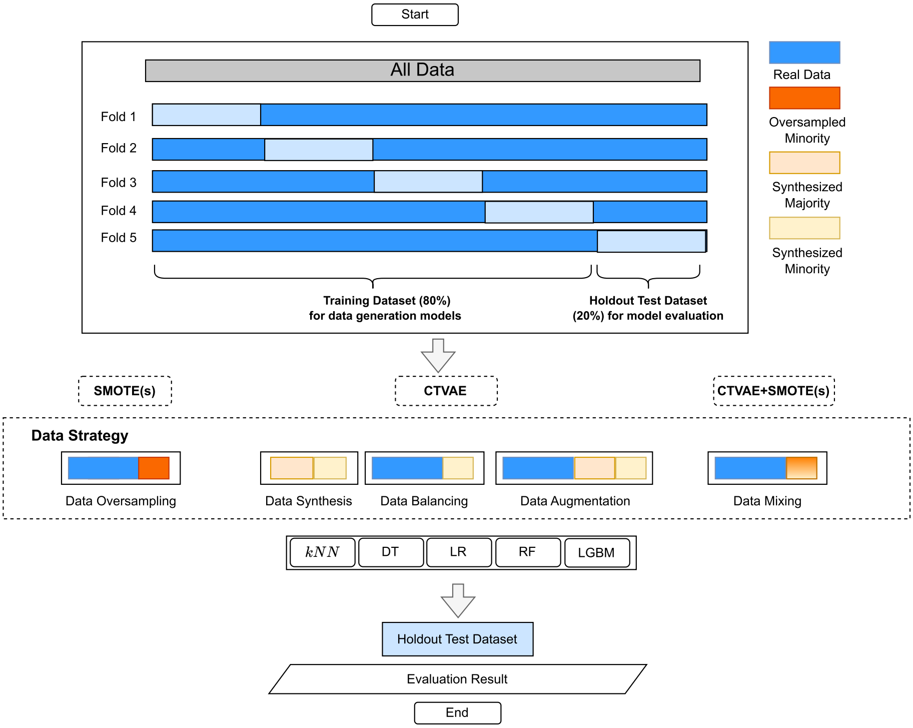

Fig. 4 Our experimental framework

1. **Baselines (Data_Res)**: Models are trained on the original training data using traditional resampling techniques and evaluated on real test data. This serves as a baseline with traditional **Data Resampling** methods like SMOTE and its variants.
2. **Data Synthesis (Data_Syn)**: Models are trained exclusively on synthetic data generated by CTVAE, with both majority and minority samples synthesized. This approach evaluates how well models trained only on synthetic data perform when tested on real-world test data.
3. **Data Augmentation (Data_Aug)**: Models are trained on real training data augmented with additional synthetic samples generated by CTVAE. Both majority and minority samples are synthesized, expanding the dataset while preserving the original class distribution.
4. **Data Balancing (Data_Bal)**: Synthetic samples are used specifically to balance the dataset by generating additional minority class instances, thus addressing the class imbalance. This strategy evaluates the effect of balancing the dataset through targeted synthetic sampling.
5. **Data Mixing (Data_Mix)**: Ensemble models are trained on real data, balanced through a combination of synthetic minority samples from CTVAE and other oversampling techniques, with equal proportions between CTVAE and traditional methods. To indicate this integration, a "+" sign is added to the model's name; for example, SMOTE combined with CTVAE is referred to as SMOTE+.

4.3 Evaluation settings

To assess the robustness of the proposed algorithms comprehensively, we employed five widely recognized classification algorithms: k-Nearest Neighbors (k-NN), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), and Light Gradient Boosting Machine (LightGBM). This selection was made to represent various algorithmic perspectives, covering a spectrum of model complexities from simple linear models to sophisticated ensemble learning models [38]. Additionally, these algorithms exhibit varying sensitivities to class imbalances. Thus, our objective was to thoroughly evaluate the robustness of the proposed algorithms across diverse scenarios, demonstrating their efficacy in enhancing classification performance while remaining resilient to different classification algorithms [36]. To ensure a fair comparison, we maintained consistent default hyperparameters for each classification algorithm [32]. CTVAE, as a deep learning-based model, is naturally slower than the baseline traditional models. To ensure a fair comparison, we used the same hyperparameters across all datasets, following the method of previous studies [39, 40]. Specifically, we set the learning rates for the encoder and decoder (α_{enc} and α_{dec}) to 0.0001 and the contrastive loss coefficient (β) to 0.001. This consistent setup helped keep the evaluation fair and uniform.

We selected Precision–Recall Area Under the Curve (PRAUC) as the primary evaluation metric because it is particularly effective for assessing model performance in class-imbalanced scenarios [41]. To provide a more comprehensive evaluation, we also reported the Receiver Operating Characteristic Area Under the Curve (ROCAUC). PRAUC emphasizes precision and recall, offering a focused view of the model's effectiveness in correctly identifying the minority class, whereas ROCAUC evaluates the model's overall ability to distinguish between classes. The inclusion of both metrics ensures a well-rounded and balanced assessment of the classification algorithms [42].

To validate the effectiveness of integrating CTVAE with traditional resampling techniques, we applied the 5x2 cross-validation (5x2cv) paired t-test, a robust statistical method designed for model comparison [43]. This test involves performing five iterations of twofold cross-validation, where in each iteration, the dataset is randomly split into two equal parts. Each model is alternately trained and tested on these folds, yielding paired performance differences. The test statistic is calculated as:

$$t = \frac{\bar{d}}{\sqrt{\frac{\sum_{i=1}^{5}\sum_{j=1}^{2}(d_{ij}-\bar{d})^2}{10}}} \tag{1}$$

where $\bar{d}$ is the mean difference in performance and d_{ij} represents the difference for fold j in iteration i. This approach provides reliable comparisons by mitigating biases that may arise from single data splits, ensuring the validity of our analysis. All experiments were conducted on a server equipped with an Intel Xeon CPU and a GPU with 16GB of memory, ensuring sufficient computational resources for our analyses.

5 Results and discussion

5.1 Comparative study across diverse datasets

Table 2 presents the average PRAUC and ROCAUC values, along with their rankings in parentheses, across 16 datasets and 5 data strategies. The SMOTEENN+ (ensemble of CTVAE

Table 2 Average PRAUC and ROCAUC across 16 datasets, where higher means better

PRAUC	Baselines				Data_Syn	Data_Aug	Data_Bal	Data_Mix		
Dataset	No sampler	SMOTE	SyMProD	SMOTEENN	CTVAE	CTVAE	CTVAE	SMOTE+	SyMProD+	SMOTEENN+
EC	0.513 (8)	0.562 (2)	0.574 (1)	0.557 (3)	0.321 (10)	0.425 (9)	0.556 (4)	0.547 (7)	0.555 (5)	0.554 (6)
LM	0.810 (8)	0.858 (6)	0.887 (2)	0.865 (4)	0.282 (10)	0.486 (9)	0.818 (7)	0.860 (5)	0.891 (1)	0.869 (3)
AR	0.564 (9)	0.714 (2)	0.694 (4)	0.662 (6)	0.449 (10)	0.627 (7)	0.573 (8)	0.722 (1)	0.698 (3)	0.668 (5)
OI	0.778 (3)	0.682 (7)	0.702 (5)	0.640 (10)	0.655 (8)	0.789 (1)	0.788 (2)	0.691 (6)	0.708 (4)	0.643 (9)
UC	0.459 (7)	0.462 (6)	0.448 (9)	0.471 (5)	0.437 (10)	0.482 (3)	0.483 (2)	0.478 (4)	0.452 (8)	0.488 (1)
YM	0.114 (9)	0.139 (6)	0.141 (5)	0.192 (3)	0.104 (10)	0.127 (7)	0.115 (8)	0.196 (1)	0.141 (4)	0.194 (2)
OZ	0.150 (6)	0.128 (9)	0.138 (8)	0.182 (3)	0.099 (10)	0.166 (4)	0.152 (5)	0.205 (1)	0.140 (7)	0.184 (2)
SI	0.343 (8)	0.367 (4)	0.360 (6)	0.414 (2)	0.217 (10)	0.340 (9)	0.350 (7)	0.378 (3)	0.360 (5)	0.421 (1)
TH	0.644 (8)	0.735 (7)	0.758 (4)	0.763 (2)	0.353 (10)	0.544 (9)	0.750 (6)	0.753 (5)	0.759 (3)	0.769 (1)
AB	0.250 (9)	0.287 (6)	0.288 (5)	0.380 (2)	0.238 (10)	0.272 (7)	0.252 (8)	0.311 (3)	0.290 (4)	0.383 (1)
WQ	0.432 (2)	0.372 (8)	0.383 (6)	0.385 (4)	0.279 (10)	0.332 (9)	0.437 (1)	0.377 (7)	0.385 (5)	0.389 (3)
OD	0.934 (7)	0.935 (6)	0.941 (3)	0.930 (8)	0.888 (10)	0.899 (9)	0.940 (4)	0.941 (2)	0.942 (1)	0.938 (5)
ST	0.467 (8)	0.632 (7)	0.645 (2)	0.642 (6)	0.232 (10)	0.387 (9)	0.643 (4)	0.644 (3)	0.646 (1)	0.643 (5)
CO	0.443 (8)	0.612 (7)	0.624 (1)	0.614 (5)	0.283 (10)	0.398 (9)	0.619 (3)	0.617 (4)	0.622 (2)	0.613 (6)
PD	0.963 (1)	0.954 (6)	0.913 (10)	0.953 (7)	0.934 (9)	0.942 (8)	0.963 (1)	0.956 (4)	0.954 (5)	0.957 (3)
AD	0.749 (2)	0.732 (9)	0.733 (8)	0.734 (7)	0.210 (10)	0.742 (4)	0.739 (6)	0.753 (1)	0.743 (3)	0.741 (5)
Average	0.538 (6.4)	0.573 (6.1)	0.577 (4.9)	0.586 (4.8)	0.374 (9.8)	0.497 (7.1)	0.574 (4.8)	0.589 (3.6)	0.580 (3.8)	0.591 (3.6)

Table 2 continued

PRAUC	Baselines				Data_Syn	Data_Aug	Data_Bal	Data_Mix		
Dataset	No sampler	SMOTE	SyMProD	SMOTEENN	CTVAE	CTVAE	CTVAE	SMOTE+	SyMProD+	SMOTEENN+
EC	0.658 (10)	0.934 (5)	0.961 (3)	0.970 (1)	0.702 (9)	0.823 (8)	0.892 (7)	0.935 (4)	0.927 (6)	0.963 (2)
LM	0.887 (8)	0.934 (6)	0.961 (4)	0.970 (2)	0.702 (10)	0.823 (9)	0.892 (7)	0.936 (5)	0.964 (3)	0.975 (1)
AR	0.901 (9)	0.960 (4)	0.963 (3)	0.920 (6)	0.836 (10)	0.919 (7)	0.910 (8)	0.969 (2)	0.970 (1)	0.925 (5)
OI	0.926 (5)	0.921 (7)	0.941 (2)	0.913 (9)	0.874 (10)	0.920 (8)	0.937 (3)	0.930 (4)	0.943 (1)	0.922 (6)
UC	0.808 (9)	0.830 (5)	0.829 (6)	0.815 (8)	0.791 (10)	0.816 (7)	0.855 (2)	0.837 (4)	0.838 (3)	0.865 (1)
YM	0.569 (9)	0.576 (6)	0.585 (5)	0.575 (7)	0.560 (10)	0.602 (2)	0.575 (7)	0.613 (1)	0.596 (3)	0.595 (4)
OZ	0.726 (7)	0.700 (9)	0.731 (6)	0.758 (3)	0.670 (10)	0.704 (8)	0.732 (5)	0.776 (1)	0.735 (4)	0.769 (2)
SI	0.817 (7)	0.817 (7)	0.820 (6)	0.841 (2)	0.358 (10)	0.807 (9)	0.832 (4)	0.835 (3)	0.832 (5)	0.855 (1)
TH	0.768 (8)	0.779 (6)	0.786 (2)	0.814 (1)	0.729 (10)	0.749 (9)	0.772 (7)	0.780 (5)	0.781 (4)	0.782 (3)
AB	0.768 (8)	0.779 (5)	0.786 (4)	0.814 (2)	0.729 (10)	0.749 (9)	0.772 (6)	0.771 (7)	0.799 (3)	0.822 (1)
WQ	0.831 (4)	0.820 (8)	0.828 (6)	0.828 (5)	0.776 (10)	0.813 (9)	0.838 (2)	0.827 (7)	0.838 (1)	0.836 (3)
OD	0.969 (8)	0.975 (6)	0.983 (2)	0.976 (5)	0.956 (10)	0.960 (9)	0.974 (7)	0.980 (3)	0.989 (1)	0.979 (4)
ST	0.991 (2)	0.989 (4)	0.988 (6)	0.984 (8)	0.978 (10)	0.983 (9)	0.995 (1)	0.988 (5)	0.985 (7)	0.990 (3)
CO	0.871 (1)	0.865 (7)	0.864 (8)	0.869 (4)	0.530 (10)	0.861 (9)	0.870 (2)	0.869 (3)	0.868 (6)	0.868 (5)
PD	0.991 (6)	0.992 (4)	0.988 (8)	0.992 (4)	0.978 (10)	0.983 (9)	0.995 (2)	0.996 (1)	0.989 (7)	0.995 (2)
AD	0.871 (4)	0.865 (7)	0.864 (8)	0.869 (6)	0.530 (10)	0.861 (9)	0.870 (5)	0.884 (1)	0.881 (2)	0.879 (3)
Average	0.834 (6.6)	0.859 (6.0)	0.867 (4.9)	0.869 (4.6)	0.731 (9.9)	0.836 (8.1)	0.857 (4.7)	0.870 (3.5)	0.871 (3.6)	0.876 (2.9)

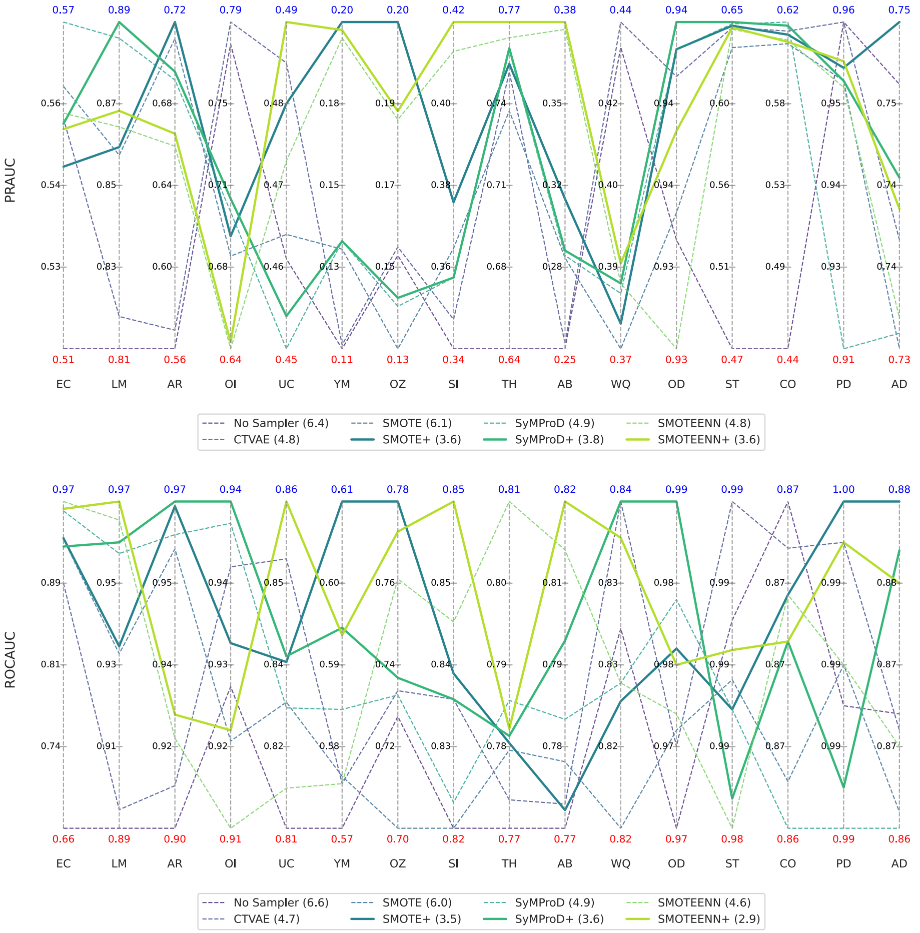

Fig. 5 Average PRAUC and ROCAUC computed w.r.t. 16 datasets, where higher indicates better. The top three models are highlighted with thicker lines

and SMOTEENN) approach within the **Data_Mix** strategy consistently runs top, achieving averages of 0.591 (3.6) and 0.876 (2.9). Similarly, SMOTE and SyMProD show improved performance when combined with CTVAE. These results highlight the potential of ensemble models that integrate conditional DGMs like CTVAE with shallow interpolation techniques such as SMOTE, effectively balancing novelty and noise in the generated samples. Another notable strategy is **Data_Bal**, which achieves PRAUC and ROCAUC averages of 0.574 (4.8) and 0.857 (4.7). This demonstrates the effectiveness of the proposed CTVAE as a data balancing technique compared to SMOTE and its variants, especially for high-dimensional datasets such as OI, UC, WQ, and PD. However, as expected, CTVAE's performance is limited on small datasets, where deep learning methods typically require larger data volumes. In such cases, SMOTE and its variants outperform CTVAE. Finally, the **Data_Syn** strategy, which relies solely on synthetic data, and the **Data_Aug** strategy both demonstrate poor performance, with average PRAUC and ROCAUC of 0.374 (9.8) and 0.731 (9.9), and 0.497 (7.1) and 0.836 (8.1), respectively. These findings demonstrate the versatility of CTVAE in

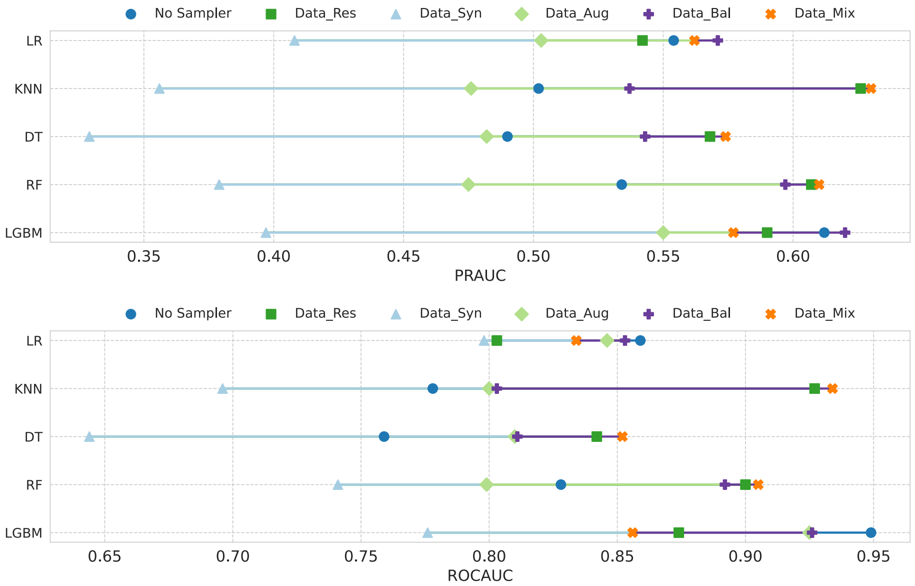

Fig. 6 Average PRAUC and AUCROC across 5 classifiers for the original data and 5 data strategies

improving classification performance on imbalanced datasets and its potential to enhance existing oversampling techniques by introducing novelty. When the appropriate data strategy is selected, CTVAE proves to be a valuable tool for addressing class imbalance challenges.

To improve readability, detailed results are also presented in Fig. 5. The **Data_Syn** and **Data_Aug** strategies are excluded due to their suboptimal performance. The parallel coordinates plot visualizes average PRAUC and ROCAUC values across 16 datasets, ordered by data size. Overall, the three ensemble models from the **Data_Mix** strategy-SMOTE+, SyMProD+, and SMOTEENN+-consistently deliver better performance. However, they underperform on datasets such as OI, OZ, and WQ, which exhibit severe imbalance ratios of 22.0, 34.0, and 26.0, respectively. In these cases, CTVAE emerges as a more effective balancing method. This is because traditional sampling techniques, like SMOTE, rely exclusively on the minority class and often struggle with extreme imbalance. In contrast, CTVAE leverages the distributions of both majority and minority classes, enhancing its robustness in such challenging scenarios. The strong performance of **Data_Mix**, alongside the comparable outcomes achieved by **Data_Bal**, highlights the importance of tailoring the strategy to the dataset's size and imbalance severity. Selecting an appropriate approach ensures optimal classification outcomes.

5.2 Comparative study across different classifiers

The average PRAUC and ROCAUC for 5 classifiers across 5 data strategies are shown in Table 3. While the **Data_Mix** approach achieves the highest overall PRAUC and ROCAUC values, it performs less effectively with LR and LGBM, where **Data_Bal** yields the best PRAUC scores. This highlights the critical need to choose data sampling methods tailored to the specific classification algorithms that are employed. For advanced classification algorithms like LGBM, the advantages of applying data balancing techniques are less pronounced.

Table 3 Average PRAUC and ROCAUC across 5 classifiers, where higher means better

PRAUC	Baselines				Data_Syn	Data_Aug	Data_Bal	Data_Mix		
Classifier	No sampler	SMOTE	SyMProD	SMOTEENN	CTVAE	CTVAE	CTVAE	SMOTE+	SyMProD+	SMOTEENN+
LR	0.554 (6)	0.555 (5)	0.545 (7)	0.542 (8)	0.408 (10)	0.503 (9)	0.571 (1)	0.569 (2)	0.565 (3)	0.562 (4)
KNN	0.502 (8)	0.573 (6)	0.575 (5)	0.626 (2)	0.356 (10)	0.476 (9)	0.537 (7)	0.590 (3)	0.579 (4)	0.630 (1)
DT	0.490 (8)	0.510 (7)	0.553 (3)	0.568 (2)	0.329 (10)	0.482 (9)	0.543 (6)	0.545 (5)	0.546 (4)	0.574 (1)
RF	0.534 (8)	0.617 (2)	0.609 (5)	0.607 (6)	0.379 (10)	0.475 (9)	0.597 (7)	0.636 (1)	0.610 (4)	0.610 (3)
LGBM	0.612 (2)	0.611 (3)	0.602 (5)	0.590 (7)	0.397 (10)	0.550 (9)	0.620 (1)	0.606 (4)	0.602 (6)	0.577 (8)
Average	0.538 (6.4)	0.573 (4.6)	0.578 (5.0)	0.587 (5.0)	0.374 (10.0)	0.497 (9.0)	0.574 (4.4)	0.588 (3.0)	0.578 (4.2)	0.591 (3.4)

ROCAUC	Baselines				Data_Syn	Data_Aug	Data_Bal	Data_Mix		
Classifier	No sampler	SMOTE	SyMProD	SMOTEENN	CTVAE	CTVAE	CTVAE	SMOTE+	SyMProD+	SMOTEENN+
LR	0.859 (1)	0.831 (7)	0.819 (8)	0.803 (9)	0.798 (10)	0.846 (4)	0.853 (2)	0.840 (5)	0.848 (3)	0.834 (6)
KNN	0.778 (9)	0.858 (6)	0.865 (5)	0.927 (2)	0.696 (10)	0.800 (8)	0.803 (7)	0.872 (3)	0.869 (4)	0.934 (1)
DT	0.759 (9)	0.763 (8)	0.831 (3)	0.842 (2)	0.644 (10)	0.810 (6)	0.811 (5)	0.805 (7)	0.819 (4)	0.852 (1)
RF	0.828 (8)	0.925 (2)	0.916 (3)	0.900 (6)	0.741 (10)	0.799 (9)	0.892 (7)	0.940 (1)	0.915 (4)	0.905 (5)
LGBM	0.949 (1)	0.916 (4)	0.905 (5)	0.874 (8)	0.776 (10)	0.925 (3)	0.926 (2)	0.895 (7)	0.903 (6)	0.856 (9)
Average	0.834 (5.6)	0.859 (5.4)	0.867 (4.8)	0.869 (5.4)	0.731 (10.0)	0.836 (6.0)	0.857 (4.6)	0.870 (4.6)	0.871 (4.2)	0.876 (4.4)

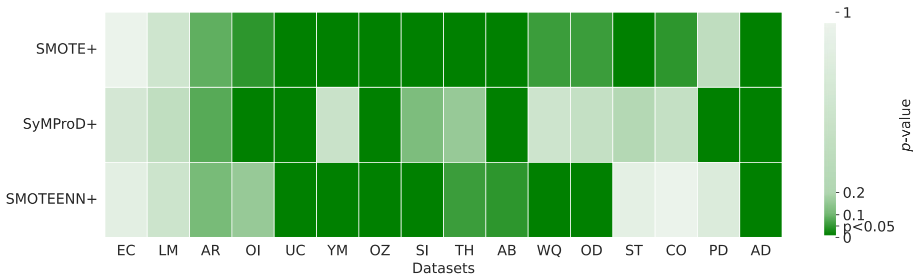

Fig. 7 Heatmap of 5x2cv paired t-test showing statistical significance ($p < 0.05$) of performance improvements through the data mixing strategy across 16 datasets for the three data sampling techniques in our baseline evaluation

For instance, with real data, LGBM achieves a PRAUC of 0.612, but the performance declines when SMOTE and its variants are applied. This suggests that data balancing techniques may not always enhance performance for more sophisticated models. On the other hand, simpler algorithms like RF benefit significantly from data-centric approaches. For example, the application of data balancing techniques improves the PRAUC from 0.534 to a peak of 0.636. These observations are further illustrated in Fig. 6, which demonstrates that while data balancing techniques are beneficial for simpler classification methods, their impact on advanced models may be limited. Notably, **Data_Bal** with CTVAE stands out as the only strategy that improves the performance of LGBM in terms of PRAUC. Therefore, it is crucial to carefully select and apply data sampling strategies based on the characteristics and requirements of the classification algorithms in use.

5.3 Statistical significance analysis

To validate our hypothesis about the effectiveness of CTVAE in enhancing traditional resampling techniques by generating novel minority class samples, we conducted a statistical significance analysis across all 16 datasets. Figure 7 illustrates the results of the 5x2cv paired t-test, where the intensity of the green shading represents the p-values, with darker shades indicating lower p-values and higher statistical significance. The analysis shows that integrating CTVAE significantly improves the performance of traditional resampling techniques, particularly for SMOTE and larger datasets ($p < 0.05$). However, the improvement is less pronounced for SyMProD and SMOTEENN, likely due to the complexity of these algorithms, which already include advanced noise filtering mechanisms. These findings highlight the potential of combining conditional DGMs like CTVAE with traditional resampling methods to create robust ensemble models, offering valuable insights for developing future resampling strategies that address class imbalance challenges effectively.

6 Conclusion

In this paper, CTVAE is introduced as a novel method for addressing class imbalance in tabular data using conditional Variational Autoencoders enhanced with contrastive learning. The approach includes the development of a contrastive oversampling model, which leverages contrastive learning to refine the latent space representation, ensuring that the generated samples better capture the complexities of the minority class distribution. Additionally, we propose an ensemble framework that integrates CTVAE with various SMOTE variants, further enhancing its ability to tackle class imbalance. Our experimental results demonstrate that CTVAE effectively mitigates class imbalance, leading to improvements in classification performance across multiple datasets. These findings highlight CTVAE's potential to contribute to the field of imbalanced learning, offering a robust and versatile solution that outperforms traditional oversampling techniques. Future research directions may explore extending CTVAE to diverse domains and datasets with complex structures, broadening its applicability beyond tabular data scenarios while ensuring scalability and robust performance.

Acknowledgements This research was supported by Ho Chi Minh City Open University under the Grant Number E2024.03.1.

Author Contributions AXW involved in conceptualization, data curation, methodology, software, investigation, formal analysis, visualization, writing—original draft. MQL took part in conceptualization, investigation, formal analysis, validation, funding acquisition, writing—review and editing. H-TD carried out investigation, writing—review and editing. BNV took part in investigation, writing—review and editing. BPN carried out conceptualization, methodology, investigation, validation, funding acquisition, supervision, writing—review and editing.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data Availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

References

1. Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6(1):1–54
2. Dongdong L, Ziqiu C, Bolu W, Zhe W, Hai Y, Wenli D (2021) Entropy-based hybrid sampling ensemble learning for imbalanced data. Int J Intell Syst 36(7):3039–3067
3. Rehman AU, Butt WH, Ali TM, Javaid S, Almufareh MF, Humayun M, Rahman H, Mir A, Shaheen M (2024) A machine learning-based framework for accurate and early diagnosis of liver diseases: a comprehensive study on feature selection, data imbalance, and algorithmic performance. Int J Intell Syst 2024(1):6111312
4. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
5. Dablain D, Krawczyk B, Chawla NV (2022) DeepSMOTE: fusing deep learning and SMOTE for imbalanced data. IEEE Trans Neural Netw Learn Syst 34(9):6390–6404
6. Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G (2022) Deep neural networks and tabular data: a survey. IEEE Trans Neural Netw Learn Syst 35(6):7499–7519
7. Yang X, Ye T, Yuan X, Zhu W, Mei X, Zhou F (2024) A novel data augmentation method based on denoising diffusion probabilistic model for fault diagnosis under imbalanced data. IEEE Trans Ind Inf 5:7820–7831

8. Wang AX, Chukova SS, Simpson CR, Nguyen BP (2024) Challenges and opportunities of generative models on tabular data. Appl Soft Comput 166:112223
9. Vivekananthan S (2024) Comparative analysis of generative models: enhancing image synthesis with VAEs, GANs, and stable diffusion. arXiv:2408.08751
10. Bai J, Kong S, Gomes CP (2022) Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In: International conference on machine learning. PMLR, pp 1383–1398
11. Wang Y, Zhang H, Liu Z, Yang L, Yu PS (2022) Contrastvae: contrastive variational autoencoder for sequential recommendation. In: Proceedings of the 31st ACM international conference on information and knowledge management, pp 2056–2066
12. Wang AX, Chukova SS, Nguyen BP (2023) Synthetic minority oversampling using edited displacement-based k-nearest neighbors. Appl Soft Comput 148:110895
13. Valdivia A, Sánchez-Monedero J, Casillas J (2021) How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. Int J Intell Syst 36(4):1619–1643
14. Sonoda R (2023) Fair oversampling technique using heterogeneous clusters. Inf Sci 640:119059
15. Huang C, Li Y, Loy CC, Tang X (2019) Deep imbalanced learning for face recognition and attribute prediction. IEEE Trans Pattern Anal Mach Intell 42(11):2781–2794
16. Yang K, Yu Z, Chen W, Liang Z, Chen CP (2024) Solving the imbalanced problem by metric learning and oversampling. IEEE Trans Knowl Data Eng 12:9294–9307
17. Lin M, Yang K, Yu Z, Shi Y, Chen CP (2023) Hybrid ensemble broad learning system for network intrusion detection. IEEE Trans Ind Inf 4:5622–5633
18. Chen W, Yang K, Yu Z, Shi Y, Chen C (2024) A survey on imbalanced learning: latest research, applications and future directions. Artif Intell Rev 57(6):1–51
19. Wang AX, Chukova SS, Sporle A, Milne BJ, Simpson CR, Nguyen BP (2024) Enhancing public research on citizen data: an empirical investigation of data synthesis using Statistics New Zealand's integrated data infrastructure. Inf Process Manag 61(1):103558
20. Nguyen HM, Cooper EW, Kamei K (2009) Borderline over-sampling for imbalanced data classification. In: Proceedings of the fifth international workshop on computational intelligence and applications. IEEE, pp 24–29
21. He H, Bai Y, Garcia EA, Li S (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the 2008 IEEE international joint conference on neural networks. IEEE, pp 1322–1328
22. Kunakorntum I, Hinthong W, Phunchongharn P (2020) A synthetic minority based on probabilistic distribution (SyMProD) oversampling for imbalanced datasets. IEEE Access 8:114692–114704
23. Batista GE, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl 6(1):20–29
24. Kim J, Kong J, Son J (2021) Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: International conference on machine learning. PMLR, pp 5530–5540
25. Damm S, Forster D, Velychko D, Dai Z, Fischer A, Lücke J (2023) The ELBO of variational autoencoders converges to a sum of entropies. In: International conference on artificial intelligence and statistics. PMLR, pp 3931–3960
26. Zheng Y, He T, Qiu Y, Wipf DP (2022) Learning manifold dimensions with conditional variational autoencoders. Adv Neural Inf Process Syst 35:34709
27. Aneja J, Schwing A, Kautz J, Vahdat A (2021) A contrastive learning approach for training variational autoencoder priors. Adv Neural Inf Process Syst 34:480–493
28. Xie Z, Liu C, Zhang Y, Lu H, Wang D, Ding Y (2021) Adversarial and contrastive variational autoencoder for sequential recommendation. In: Proceedings of the web conference 2021, pp 449–459
29. Hu M-f, Liu Z-Y, Liu J-W (2024) mcVAE: disentangling by mean constraint. Vis Comput 40(2):1229–1243
30. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K (2019) Modeling tabular data using conditional GAN. In: Advances in neural information processing systems, pp 7335–7345
31. Nasios N, Bors AG (2006) Variational learning for Gaussian mixture models. IEEE Trans Syst Man Cybern Part B (Cybernetics) 36(4):849–862
32. Wang AX, Nguyen BP (2025) TTVAE: transformer-based generative modeling for tabular data generation. Artif Intell 340:104292
33. Wang AX, Nguyen BP (2025) Deterministic Autoencoder using Wasserstein loss for tabular data generation. Neural Netw 185:107208. https://doi.org/10.1016/j.neunet.2025.107208
34. Gorishniy Y, Rubachev I, Kartashev N, Shlenskii D, Kotelnikov A, Babenko H ArtemZhang, Zhang J, Srinivasan B, Shen Z, Qin X, Faloutsos C, Rangwala H, Karypis G (2024) Mixed-type tabular data synthesis with score-based diffusion in latent space. In: The 12th international conference on learning representations (ICLR). https://openreview.net/pdf?id=4Ay23yeuz0

35. Borisov V, Seßler K, Leemann T, Pawelczyk M, Kasneci G (2023) Language models are realistic tabular data generators. In: International conference on learning representations, pp 1–18
36. Wang AX, Simpson CR, Nguyen BP (2025) Blending is all you need: data-centric ensemble synthetic data. Inf Sci 691:121610
37. Azhar NA, Pozi MSM, Din AM, Jatowt A (2022) An investigation of SMOTE based methods for imbalanced datasets with data complexity analysis. IEEE Trans Knowl Data Eng 35(7):6651–6672
38. Wang AX, Chukova SS, Nguyen BP (2023) Data-centric AI to improve churn prediction with synthetic data. In: The 3rd international conference on computer, control and robotics (ICCCR 2023). IEEE, pp 409–413
39. Zhao Z, Kunar A, Birke R, Scheer H, Chen LY (2024) CTAB-GAN+: enhancing tabular data synthesis. Front Big Data. https://doi.org/10.3389/fdata.2023.1296508
40. Kotelnikov A, Baranchuk D, Rubachev I, Babenko A (2023) TABDDPM: modelling tabular data with diffusion models. In: International conference on machine learning, p 17564
41. Wang AX, Chukova SS, Simpson CR, Nguyen BP (2023) Data-centric AI to improve early detection of mental illness. In: The 2023 IEEE statistical signal processing workshop (SSP 2023). IEEE, pp 369–373
42. Kim M, Hwang K-B (2022) An empirical evaluation of sampling methods for the classification of imbalanced data. PLoS ONE 17(7):0271260
43. Wang AX, Chukova SS, Nguyen BP (2022) Implementation and analysis of centroid displacement-based k-nearest neighbors. In: International conference on advanced data mining and applications. Springer, Berlin, pp 431–443

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Alex X. Wang is a data scientist and statistician specializing in deep learning, machine learning, and AI applications for health care. He earned both his M.Sc. and Ph.D. in Data Science from Victoria University of Wellington, New Zealand. Currently, he leads the Data Science team at the Ministry of Business, Innovation and Employment, focusing on data-centric AI, tabular data synthesis, and explainable AI in social data applications. Dr. Wang has authored over 20 publications in machine learning and AI applied to health care, contributing to the field.

Minh Quang Le obtained a Ph.D. degree in Mathematics from University at Buffalo, USA. His research interests lie in the intersection of computational topology, geometry, and deep neural networks.

Huu-Thanh Duong holds a Master's degree in Information Systems from the University of Science, Vietnam National University, Ho Chi Minh City. His primary research interests lie in natural language processing and deep learning, where he is passionate about exploring innovative techniques to enhance machine understanding of human language. Throughout his academic journey, he has dedicated his efforts to advancing the field and contributing to the development of intelligent systems.

Bay Nguyen Van received his master's degree from Ho Chi Minh City Open University, Vietnam. He is a lecturer at the same university, with research interests in machine learning and optical character recognition.

Binh P. Nguyen holds a Ph.D. in Electrical and Computer Engineering from the National University of Singapore. Currently, he is a Senior Lecturer in Data Science at the School of Mathematics and Statistics, Victoria University of Wellington (VUW), New Zealand. Throughout his career, he has received multiple awards for his contributions to developing innovative medical devices and applying visual analytics and data science to advance health care. Since joining VUW in 2018, Dr. Nguyen has been at the forefront of innovating artificial intelligence (AI) for health informatics, bioinformatics, and drug discovery. His research group is one of the few, if not the only, in New Zealand focused on developing AI-driven methods for drug discovery. Dr. Nguyen has published over 130 papers in international journals and conference proceedings.